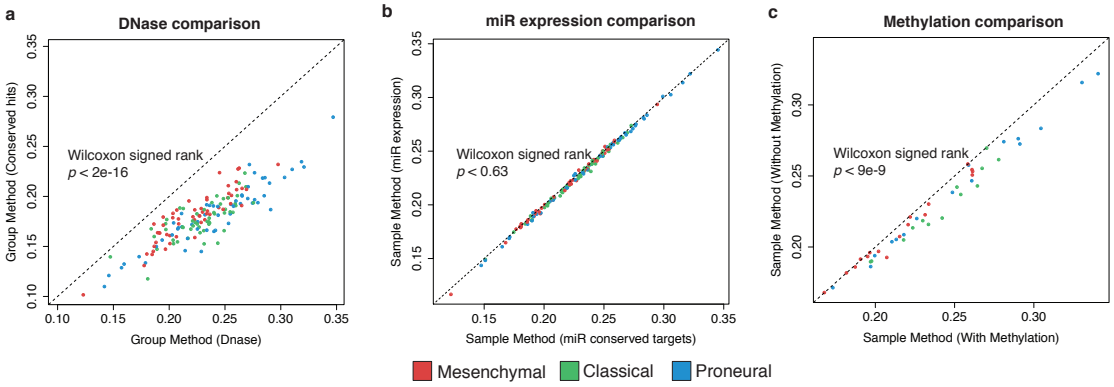
Supplementary material: Inferring transcriptional and microRNA-mediated regulatory programs in glioblastoma

Manu Setty, Karim Helmy, Aly A. Khan, Joachim Silber, Aaron Arvey, Frank Neezen, Phaedra Agius, Jason T. Huse, Eric C. Holland, and Christina S. Leslie

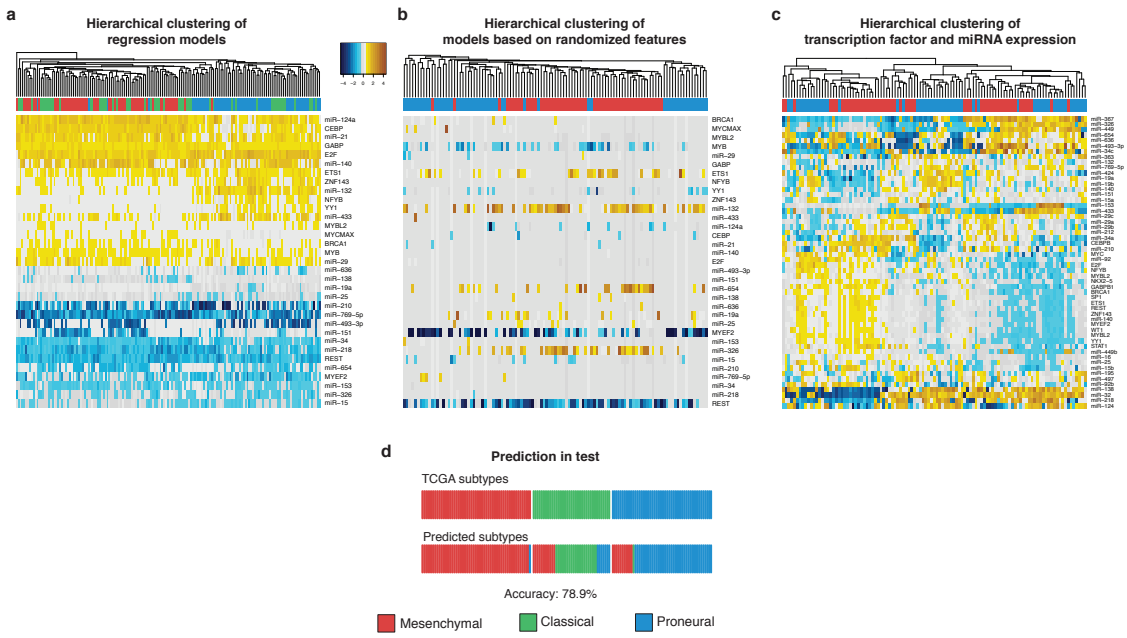
Table of Contents

Supplementary Figures



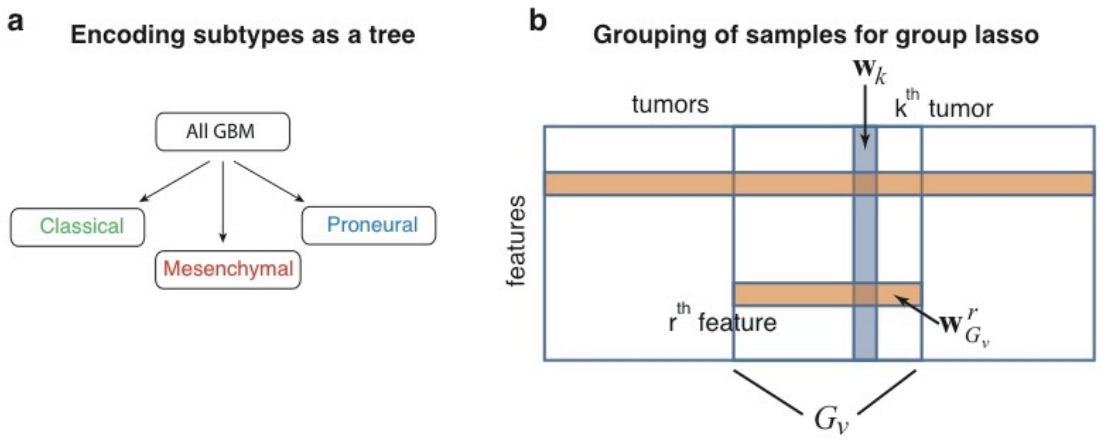
Supplementary Figure 1. Use of DNase-seq and DNA methylation data improves model accuracy, but use of miRNA expression instead of conserved hits gives similar prediction performance.

(a) Using DNaseI HS data from H54 glioblastoma cell line to filter TF binding site predictions instead filtering by conservation significantly improves Spearman correlations on held-out genes (p < 2e-16). (b) Using miRNA expression levels instead of counts of conserved seeds as gene-specific miRNA features does not significantly alter the cross-validation results (p < 0.38, Wilcoxon signed rank test). (c) Including DNA methylation as a covariate significantly improves Spearman correlations of held out genes in the sample-method (p < 9e-9, Wilcoxon signed rank test).



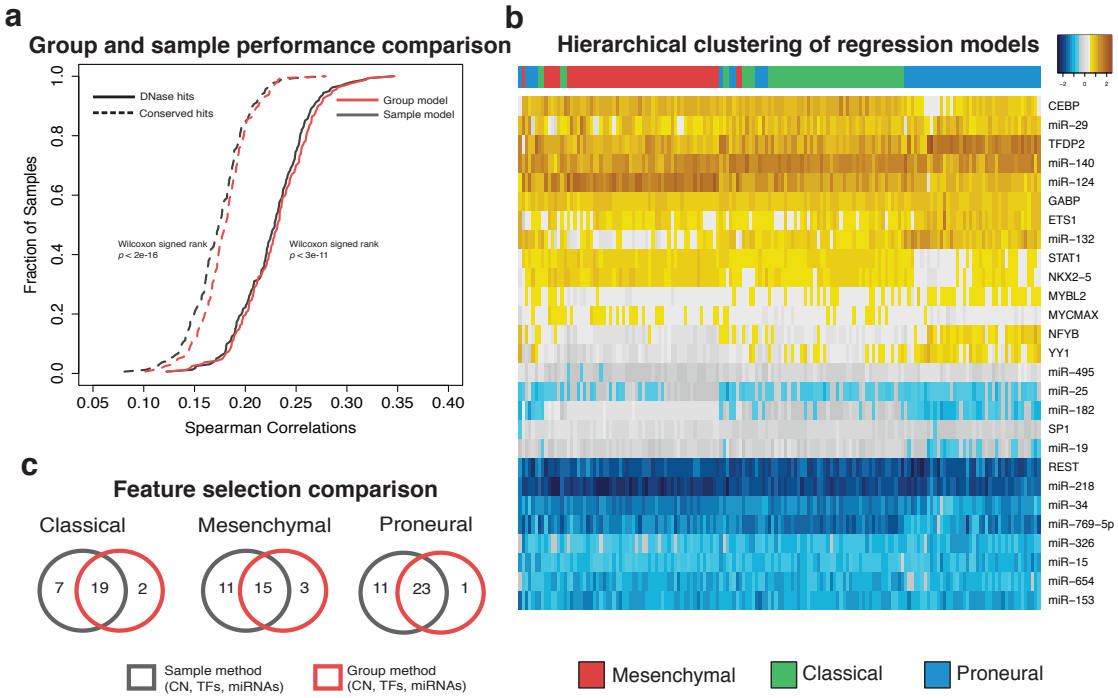
Supplementary Figure 2. Hierarchical clustering of regression models.

(a) Unsupervised hierarchical clustering of sample-by-sample regression models separates proneural samples from mesenchymal samples whereas classical samples are distributed between the two clusters. (b) Clustering of all samples based on models learned by randomized features does not distinguish the proneural and mesenchymal subtypes to the same degree and selects features with fewer targets (data not shown). (c) Clustering of all samples based on expression levels of transcription factors and microRNAs selected in the sample model again does not distinguish the proneural and mesenchymal subtypes. (d) To assess how well the group lasso models capture subtype-specific expression changes, the average subtype-specific regression models were used for subtype prediction on 160 test samples. Each test sample was assigned a subtype based on the model that best explained its expression changes relative to normal samples. This procedure achieved a classification accuracy of 78.9% on a test set of additional TCGA samples.



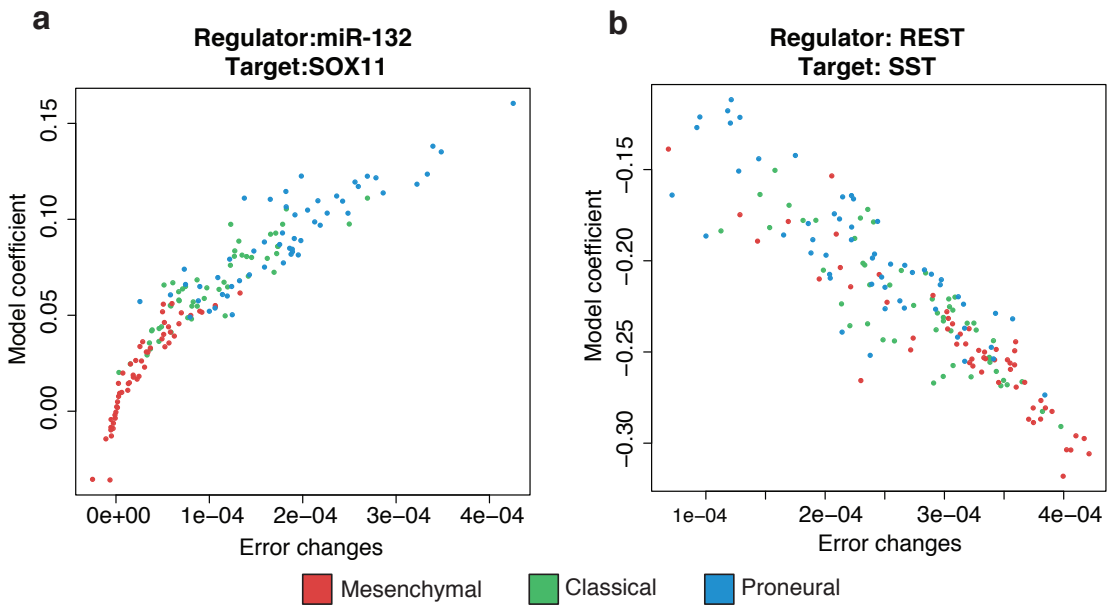
Supplementary Figure 3. Encoding of GBM subtypes as a tree in the group lasso method.

(a) Tree-guided group lasso is used for jointly learning models for all samples. Samples are represented in a hierarchy of groups based on their subtype. The root node represents all the GBM samples and lead nodes represent samples of the three subtypes. (b) Regression coefficients corresponding to a regulator are grouped across all samples and samples in a subtype as an L_2 constraint.



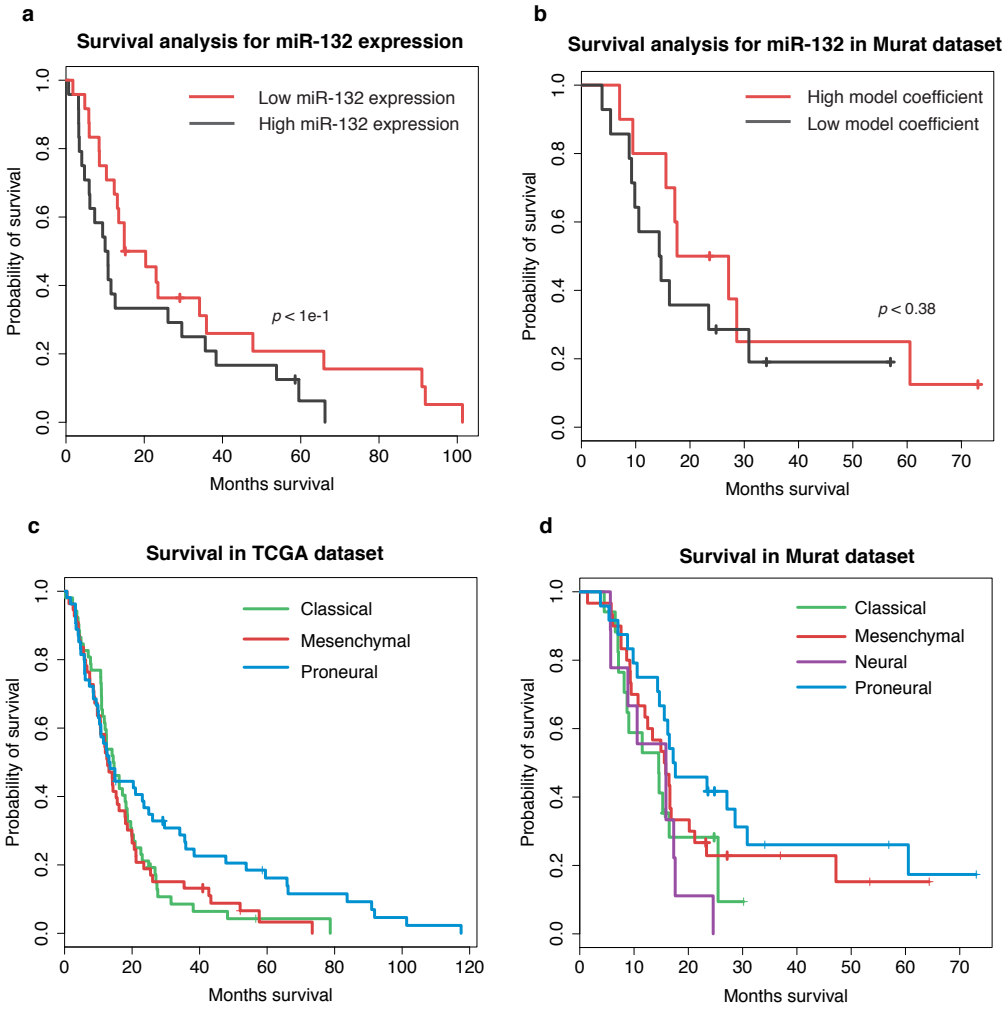
Supplementary Figure 4. Group lasso method outperforms sample-by-sample method and identifies a smaller core set of candidate regulators.

(a) Spearman correlations between predicted and actual tumor-versus-normal gene expression changes for all samples computed using 10-fold cross-validation on held out changes. Correlations from group method are significantly better than those from sample method when TF binding site prediction is done in either conserved promoter regions or DNase hypersensitive sites. (p < 3e-11, Wilcoxon signed rank test). (b) Unsupervised hierarchical clustering of tumors by their group method coefficients separates the samples of the subtypes. (c) Group models with copy numbers, TFs and miRNAs identifies to a reduced number of regulators compared to sample-by-sample method at the same FDR threshold.



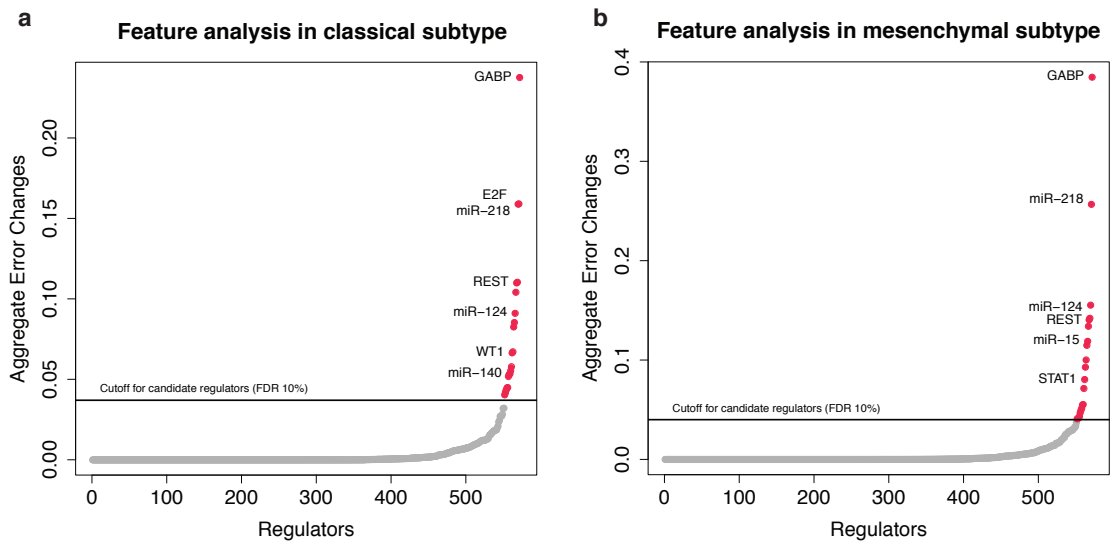
Supplementary Figure 5. Comparison of regulator-gene pair across all GBM samples.

Plot showing the variability of model coefficients of a regulator and error changes of an associated target for (a) miR-132, a proneural-specific regulator with positive regression coefficients; and (b) REST, a regulator common to all subtypes with negative regression coefficients. See also Supplementary Table 4.



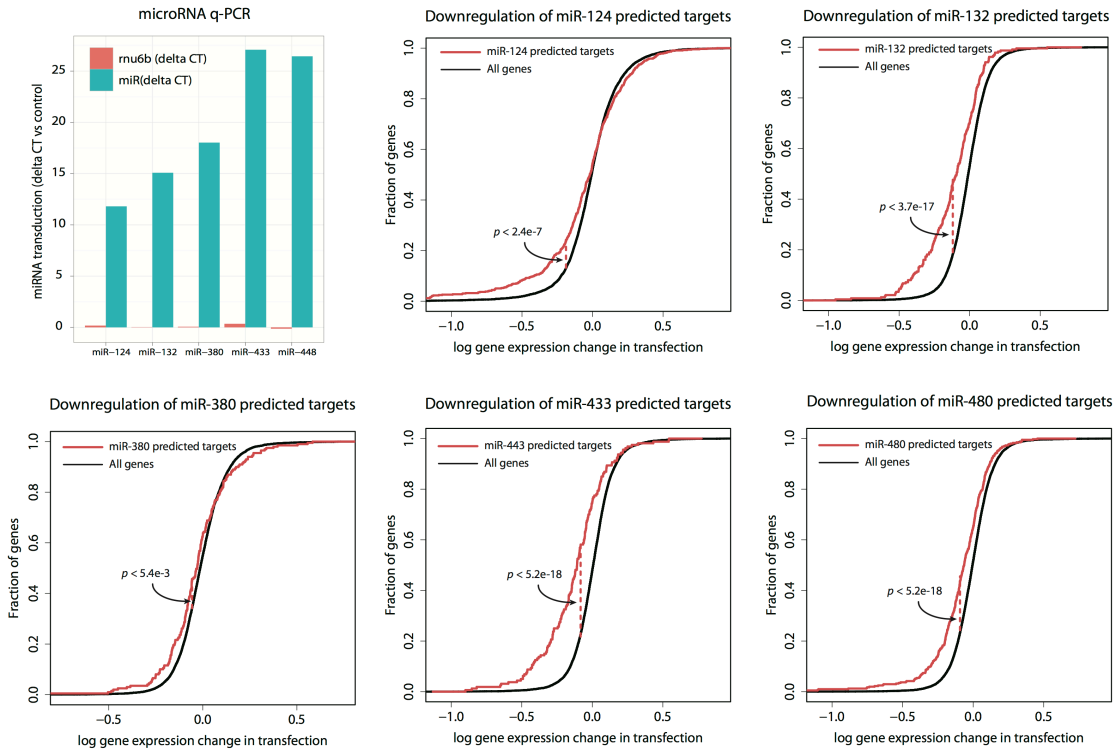
Supplementary Figure 6. Survival analysis in multiple studies.

(a) TCGA proneural patients with lower miR-132 expression (red line) show marginally better survival compared to patients with higher miR-132 expression (black line) within the proneural subtype ($p < 0.1$, log rank test). (b) Model coefficients of miR-132 in the Murat data set (Murat et al., 2008) are partially reflective of survival of the proneural patients. Patients with high model coefficient show marginally higher survival than patients with low model coefficients ($p < 0.38$, log rank test). This result is similar to the more significant survival difference in TCGA proneural patients (Figure 3c). (c) TCGA proneural patients show a significantly higher survival compared to other subtypes. (d) Predicted proneural patients from the Murat data set do not show a significantly higher survival compared to patients of other subtypes.



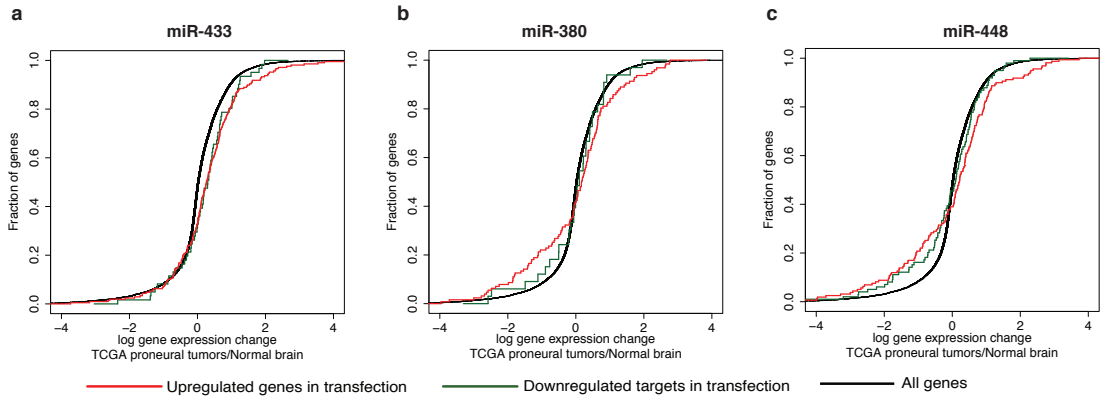
Supplementary Figure 7. Feature dependency analysis identifies regulators for the classical and mesenchymal subtypes, related to Figure 3.

Plot of aggregate error changes for (a) classical and (b) mesenchymal subtypes. Regulators are ranked based on increase in squared error across samples of a subtype after excluding the regulator from regression models. Candidate regulators are identified at an FDR of 10% relative to models trained on randomized data (Supplemental Experimental Procedures).



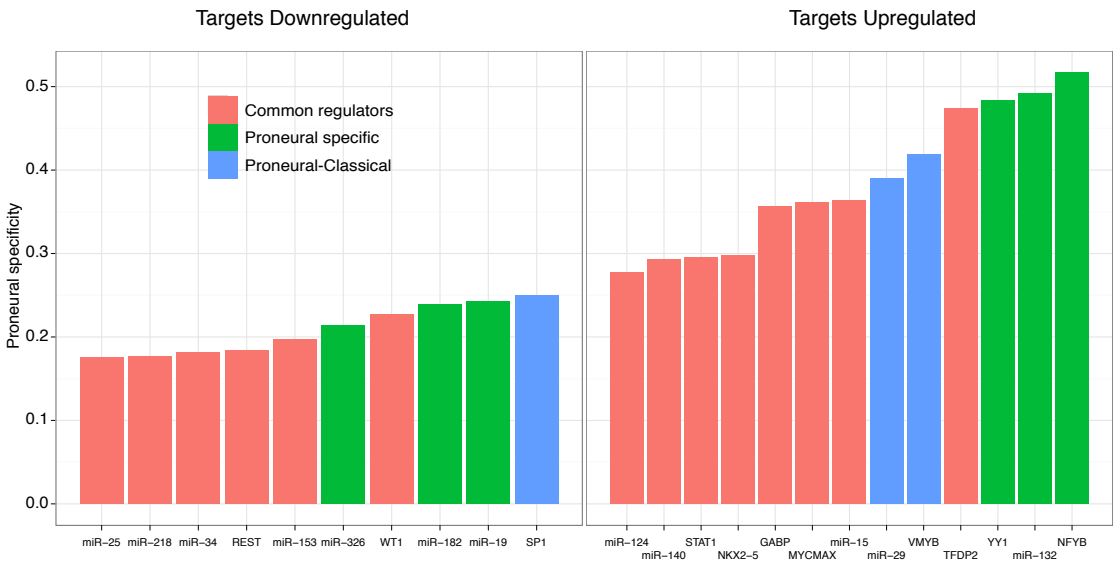
Supplementary Figure 8. Verification of microRNA transfections.

(a) qPCR using primers specific for each miRNA shows a strong induction of each miRNA immediately after transfection. (b)-(f) Predicted targets of miRNA are significantly downregulated 24 hours after transfection in each experiment ($p < 0.01$, Kolmogorov-Smirnov one-sided test).



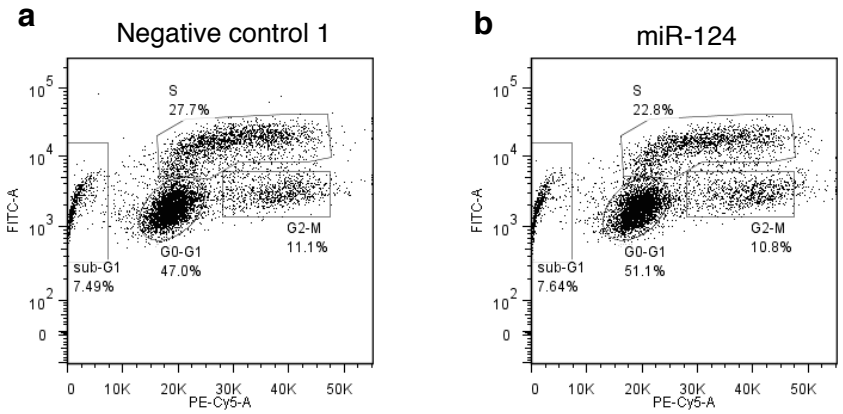
Supplementary Figure 9. Concordance plots for control microRNAs.

Expression changes after control microRNA transfections in proneural neurospheres are not significantly concordant with the expression changes in proneural tumors (FDR-corrected $p < 0.01$, Supplementary Table 6).



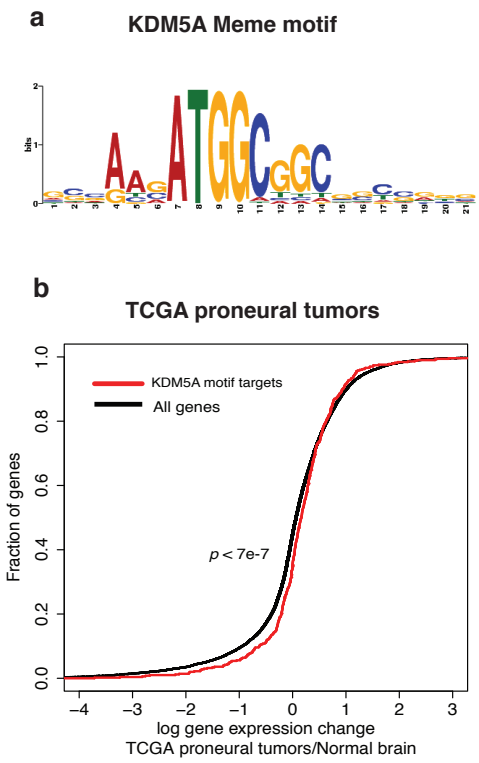
Supplementary Figure 10. Proneural-specific regulators are enriched for genes differentially expressed in proneural subtype.

The fraction of genes belonging to regulator gene sets that are significantly (left) down (a) or (right) up (b) in proneural subtype compared to other subtypes is shown. Proneural-specific regulators show higher specificity compared to the common regulators.



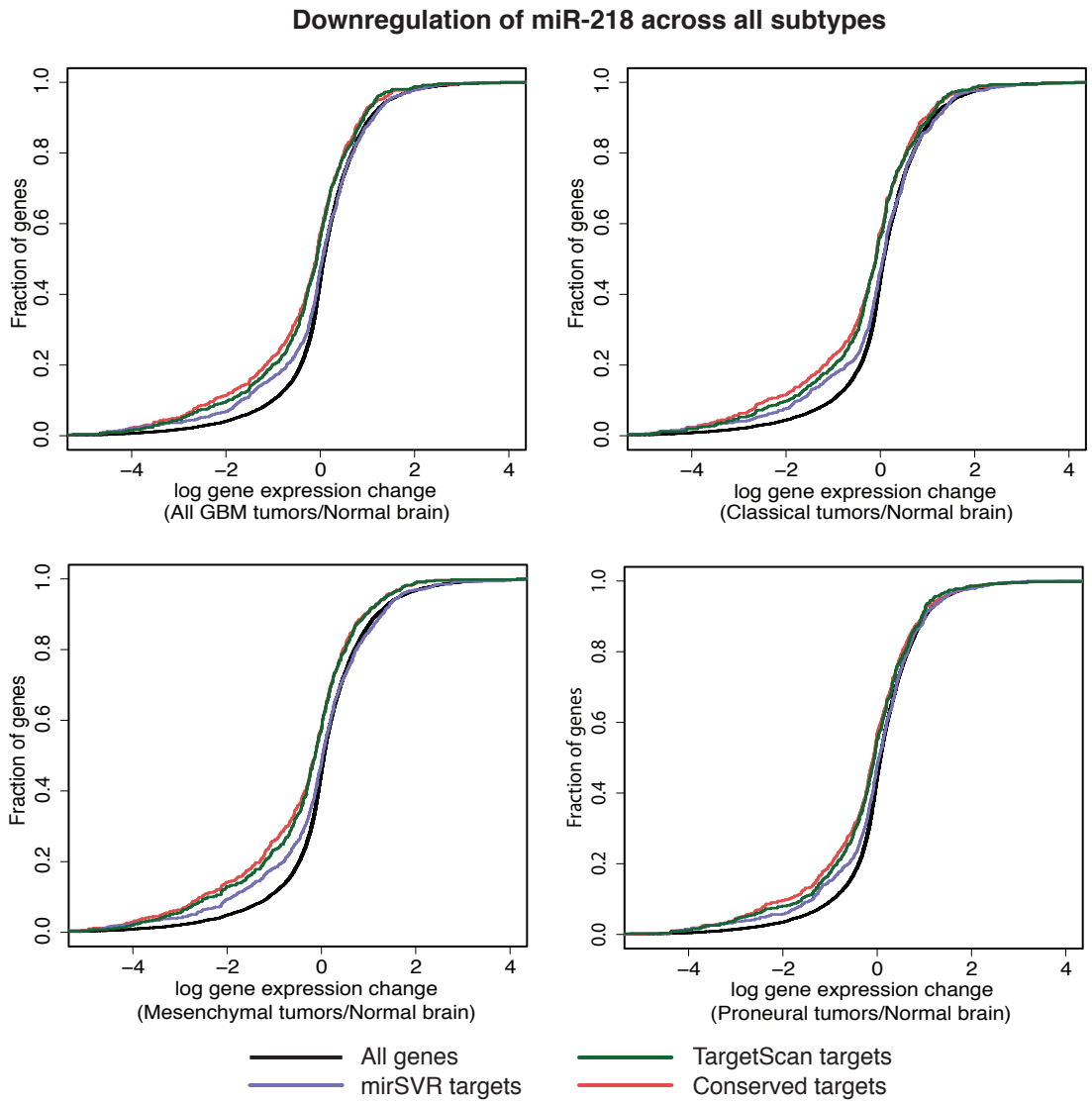
Supplementary Figure 11. Flow cytometry plots for cell proliferation assay.

Representative flow cytometry plots for the Brdu cell proliferation assay shows an increase in number of cells in G0-G1 phase in miR-124 transfection (b) compared to negative control (a).



Supplementary Figure 12. KDM5A target expression changes are concordant with miR-132 dysregulation in proneural tumors.

(a) KDM5A motif obtained from KDM5A ChIP-seq data in human embryonic stem cells. This motif was used to determine KDM5A predicted targets (Materials and methods). (b) KDM5A is a member of the miR-132 gene set in proneural tumors; consistent with this observation, KDM5A motif targets are marginally but significantly upregulated in proneural tumors (p < 7e-7, Kolmogorov-Smirnov test). In order to further assess the significance of this result, we randomly sampled the same number of genes 10,000 times and found that none of the samples gave the same significance for upregulation in proneural tumors.



Supplementary Figure 13. Predicted targets of miR-218 are downregulated in GBM, despite lower expression of this microRNA in tumor versus normal samples. The predicted targets of miR-218 are significantly downregulated across all GBM subtypes ($p < 0.01$, Kolmogorov-Smirnov test). The predicted targets were determined using three different methods: conserved 7-mer seed matches, mirSVR (Betel et al., 2010) and TargetScan (Lewis et al., 2005).

Supplementary Table Legends

Supplemenary Table 1. TCGA patient list

List of training and test TCGA patients used in the study. The subtype classification for training patients was obtained a previous study (Verhaak et al., 2010). The classification for test set was derived based on the gene signature as defined in this study (Verhaak et al., 2010) (Materials and methods).

Supplementary Table 2. TRANSFAC motifs.

Matrix identifiers of the TRANSFAC motifs used for binding site prediction. Matrices were manually curated to remove redundancy.

Supplementary Table 3. miRNA seed families.

miRNAs are grouped into families based on their 7mer seeds (positions 2-8).

Supplmentary Table 4a-c. Proneural, mesenchymal, and classical gene sets.

Target gene sets associated with regulators in each subtype as determined by feature dependency analysis (Supplemental Experimental Procedures).

Regulator	Classical	Mesenchymal	Proneural
	mean +/- SE	mean +/- SE	mean +/- SE
CEBP	0.10715 +/- 0.00245	0.10955 +/- 0.00245	0.06905 +/- 0.00315
GABP	0.08050 +/- 0.00210	0.07680 +/- 0.00240	0.07945 +/- 0.00315
MYCMAX	0.02535 +/- 0.00115	0.03160 +/- 0.00140	0.02355 +/- 0.00125
NFYB	0.01535 +/- 0.00235	-0.00430 +/- 0.00260	0.05325 +/- 0.00415
NKX2-5	0.05175 +/- 0.00175	0.06810 +/- 0.00190	0.04215 +/- 0.00175
REST	-0.22540 +/- 0.00470	-0.24770 +/- 0.00510	-0.19540 +/- 0.00530
SP1	-0.03325 +/- 0.00135	-0.02050 +/- 0.00120	-0.03565 +/- 0.00175
STAT1	0.05915 +/- 0.00145	0.06305 +/- 0.00135	0.04305 +/- 0.00185
E2F	0.12040 +/- 0.00440	0.10655 +/- 0.00425	0.15195 +/- 0.00735
MYBL2	0.03035 +/- 0.00125	0.01685 +/- 0.00135	0.04265 +/- 0.00225
WT1	-0.01955 +/- 0.00075	-0.02250 +/- 0.00080	-0.01940 +/- 0.00080
YY1	0.02260 +/- 0.00320	-0.01485 +/- 0.00285	0.04765 +/- 0.00495
miR-140	0.15290 +/- 0.00350	0.13220 +/- 0.00440	0.13310 +/- 0.00410
miR-124	0.11975 +/- 0.00255	0.18270 +/- 0.00360	0.09795 +/- 0.00305
miR-132	0.06505 +/- 0.00295	0.02195 +/- 0.00325	0.09140 +/- 0.00350
miR-153	-0.14545 +/- 0.00305	-0.16730 +/- 0.00290	-0.12750 +/- 0.00370
miR-15	-0.12300 +/- 0.00240	-0.12760 +/- 0.00260	-0.11210 +/- 0.00340
miR-182	-0.06535 +/- 0.00295	0.00000 +/- 0.00000	-0.08860 +/- 0.00340
miR-19	-0.04030 +/- 0.00300	-0.04655 +/- 0.00285	-0.06105 +/- 0.00335
miR-218	-0.23885 +/- 0.00375	-0.29435 +/- 0.00575	-0.19640 +/- 0.00470
miR-25	-0.09455 +/- 0.00315	-0.08555 +/- 0.00305	-0.08010 +/- 0.00300
miR-29	0.08710 +/- 0.00360	0.06705 +/- 0.00565	0.08070 +/- 0.00430
miR-326	-0.12685 +/- 0.00345	-0.11840 +/- 0.00330	-0.10880 +/- 0.00360
miR-34	-0.18395 +/- 0.00335	-0.20315 +/- 0.00335	-0.12935 +/- 0.00425
miR-495	-0.02690 +/- 0.00240	-0.07015 +/- 0.00265	-0.01310 +/- 0.00290
miR-654	-0.14385 +/- 0.00285	-0.16285 +/- 0.00385	-0.10125 +/- 0.00355
miR-769-5p	-0.20480 +/- 0.00390	-0.19295 +/- 0.00415	-0.13805 +/- 0.00525

Supplementary Table 5. Summary of regression model coefficients across all subtypes.

Supplementary Table 6. Murat data set patient list

List of Murat data set patients (Murat et al., 2008) used in the study for miR-132 survival analysis. The subtype classification was derived based on the gene signature defined by the TCGA study (Verhaak et al., 2010) (see Materials and methods).

Candidate regulator	Ontologies associated with gene set	Target regulation	Regulator significance *p*-value
GABP	DNA replication, DNA replication	Up	2.00E-06
TFDP2	Cell cycle process	Up	2.00E-06
miR-218	Synaptic transmission	Down	2.00E-06
miR-34	Neuron development	Down	1.00E-05
REST	Neuron death	Down	1.20E-05
miR-15	Cell morphogenesis, microtubule process	Up	1.67E-05
miR-124	Cell differentiation	Up	3.20E-05
SP1		Down	5.95E-05
miR-153	Cell-cell signaling, synaptic transmission	Down	7.16E-05
STAT1		Up	2.00E-04
WT1	Neuron development, axonogenesis	Down	2.60E-04
miR-769-5p		Down	5.20E-04
NKX2-5		Up	6.00E-04
miR-140	Cell migration	Up	8.40E-04
miR-25		Down	8.40E-04
miR-29	Matrix organization	Up	9.40E-04
CEBP		Up	1.00E-03
miR-326		Down	1.80E-03
MYCMAX	Cell cycle process	Up	2.20E-03
MYBL2	Mitotic cell cycle, chromosome organization	Up	2.40E-03
miR-654		Down	2.60E-03

Supplementary Table 7a. Gene ontologies associated with classical gene sets.

The annotations were determined based on over-represented terms for Gene Ontology "Biological Process" in gene sets.

Candidate Regulator	Ontologies associate with geneset	Target regulation	Regulator significance p-value
GABP	DNA replication, DNA replication	Up	2.00E-06
miR-218	Synaptic transmission	Down	2.00E-06
miR-124	Cell differentiation	Up	2.00E-06
REST		Down	1.75E-06
miR-34	Neuron development	Down	1.75E-06
TFDP2	Cell cycle process	Up	3.50E-06
miR-15	Cell morphogenesis, microtubule process	Up	1.60E-05
miR-153	Cell-cell signaling, synaptic transmission	Down	1.80E-05
NKX2-5		Up	5.50E-05
WT1	Neuron development, axonogenesis	Down	8.40E-05
STAT1		Up	1.80E-04
MYCMAX	Cell cycle process	Up	5.00E-04
miR-654		Down	1.46E-03
miR-769-5p		Down	1.50E-03
CEBP		Up	2.20E-03
miR-495		Down	2.20E-03
miR-25		Down	2.80E-03
miR-140	Cell migration	Up	3.40E-03

Supplementary Table 7b. Gene ontologies associated with mesenchymal gene sets.

The annotations were determined based on over-represented terms for Gene Ontology "Biological Process" in gene sets.

Regulator	Concordance *p*-value (BH corrected)
GABP	2.43E-11
TFDP2	2.64E-11
miR-218	5.13E-01
SP1	4.90E-09
miR-15	3.90E-03
REST	6.37E-01
MYBL2	1.60E-12
WT1	7.16E-03
miR-153	8.25E-01
miR-124	3.15E-05
miR-182	3.55E-03
miR-34	3.53E-01
YY1	3.06E-10
miR-29	5.34E-07
miR-140	2.49E-04
NKX2-5	2.23E-11
STAT1	8.97E-06
miR-25	4.31E-01
miR-132	1.23E-03
MYCMAX	5.16E-06
NFYB	1.12E-06
miR-326	4.09E-01
miR-19	2.28E-01

Supplementary Table 8. Gene sets for candidate proneural regulators display consistent expression changes in PDGF-driven mouse tumors.

For each predicted regulator in the GBM model, the corresponding regulated human gene set was mapped to an orthologous mouse gene set. These gene sets were assessed for up/downregulation in PDGF-driven OLIG2+ mouse tumor cells relative to mouse oligodendrocyte progenitor cells using a one-sided KS test (results pass an FDR threshold of 1% after correcting for multiple hypotheses).

	TCGA proneural tumor versus normal target/gene regulation (one-sided KS test *p*-values)					
microRNA	**Downregulation of downregulated targets in transfection**	**Upregulation of downregulated targets in transfection**	**Downregulation of upregulated genes in transfection**	**Upregulation of upregulated genes in transfection**	**Downregulation of strongly downregulated genes in transfection**	**Upregulation of strongly downregulated genes in transfection**
miR-124	1.00e+00	3.38e-08	5.29e-08	4.98e-02	1.00e+00	7.96e-10
miR-132	1.00e+00	3.65e-03	3.65e-03	4.98e-02	1.00e+00	2.14e-02
miR-433	1.00e+00	2.71e-02	9.78e-01	5.84e-06	1.00e+00	2.96e-02
miR-380	7.26e-01	9.78e-01	2.15e-02	2.49e-02	6.59e-01	1.00e+00
miR-448	6.59e-01	2.25e-01	3.44e-03	1.15e-03	9.45e-01	6.30e-01

Supplementary Table 9. Expression changes after miR-124 and miR-132 transfections are concordant with proneural tumor versus normal expression changes.

Differentially expressed predicted miRNA targets and genes from miRNA transfection experiments were assessed for up/downregulation in TCGA tumor profiles relative to normal brain tissue; *p*-values were determined using one-sided KS test with Benjamini-Hochberg correction for multiple hypothesis testing. The columns of the table that represent concordant expression changes between transfection experiments and tumor data are highlighted in red. Both miR-124 and miR-132 show concordance by the measures that (i) downregulated targets in the transfection are upregulated in proneural tumors ($p < 0.01$) and (ii) upregulated genes in the transfection are downregulated in proneural tumors ($p < 0.01$). Upregulated genes in the miR-448 transfection show significant shifts both towards upregulation and downregulation in proneural tumors, which is therefore not considered concordant. Genes that are most downregulated in miR-124 transfection are significantly upregulated in tumors. This measure shows marginal significance for miR-132 and miR-433 whereas the two control miRNAs do not show any concordance.

microRNA	Sequence
miR-380	UAUGUAAUAUGGUCCACAUCUU
miR-132	UAACAGUCUACAGCCAUGGUCG
miR-433	AUCAUGAUGGGCUCCUCGGUGU
miR-448	UUGCAUAUGUAGGAUGUCCCAU
miR-124	UAAGGCACGCGGUGAAUGCC

Supplementary Table 10. microRNA mimetic and primer sequences.

Supplementary References

Betel, D., Koppal, A., Agius, P., Sander, C., and Leslie, C. (2010). Comprehensive modeling of microRNA targets predicts functional non-conserved and non-canonical sites. Genome Biology *11*, R90.

Lewis, B.P., Burge, C.B., and Bartel, D.P. (2005). Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. Cell *120*, 15-20.

Murat, A., Migliavacca, E., Gorlia, T., Lambiv, W.L., Shay, T., Hamou, M.F., de Tribolet, N., Regli, L., Wick, W., Kouwenhoven, M.C.*, et al.* (2008). Stem cell-related "self-renewal" signature and high epidermal growth factor receptor expression associated with resistance to concomitant chemoradiotherapy in glioblastoma. J Clin Oncol *26*, 3015-3024.

Verhaak, R.G., Hoadley, K.A., Purdom, E., Wang, V., Qi, Y., Wilkerson, M.D., Miller, C.R., Ding, L., Golub, T., Mesirov, J.P.*, et al.* (2010). Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1. Cancer Cell *17*, 98-110.